Rproteomics, an implementation of the SMOS model

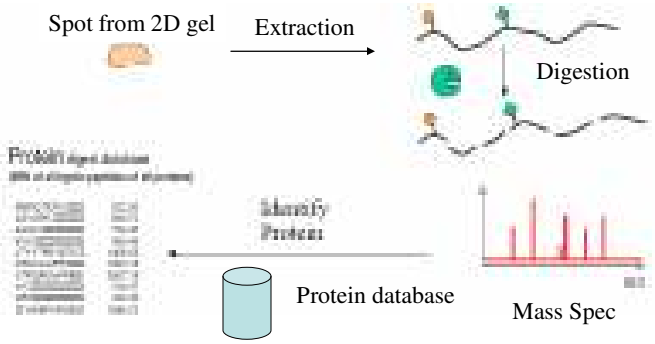
Simon Lin, Patrick McConnell,
Kim Johnson, Jennifer Shoemaker
Duke University Medical Center

Agenda

- Spectrometry-based proteomics
- SMOS
  - a proposed statistical component for MIAPE
- Rproteomics
  - an implementation of SMOS
- Engineering Plan
- Standards of data exchange in proteomics
  - PEDRo (MIAPE), PSI, and MIAME
  - EDRN network: proteomics ontology at JPL

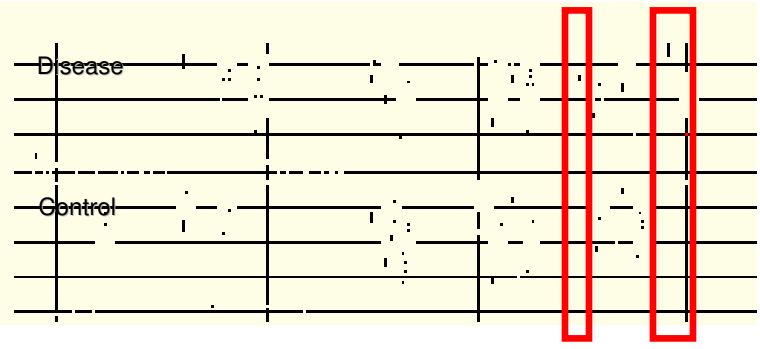
[Next: spectrometry-based]

Identification Studies



3 of 26

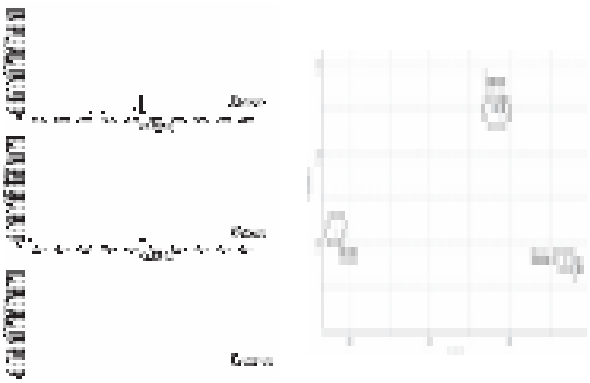
Profiling Studies



Courtesy: Ciphergen

4 of 26

Matabolomics



(Source: Bioinformatics 19: 2283, 2003)

[Next: Raw data]

Spectrum Data

- Proteomics
- Metabolomics
- HPLC, IR, etc

❶ Index j	1	2	3	4	...
❷ At j	4000.105	4000.522	4000.939	4001.356	...
❸ Yj	5518.80	5484.58	5406.03	5287.43	...

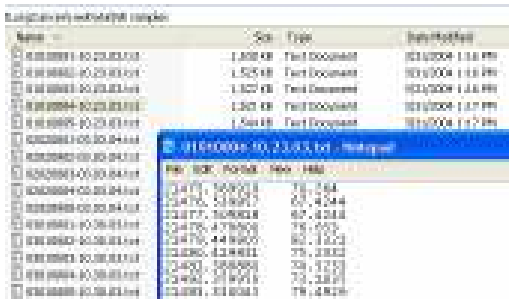
❶ Index j: also called clock tick, scan #, sample #, variable #

❷ At j: also called m/z, mass

❸ Yj: also called intensity, relative intensity, standardized intensity, abundance

Raw Data: size and format

- Size: 1.5 Mb each spectrum
- For a small study with 30 samples, 10 fractions for each sample, 10 runs for each fraction:



1.5 * 30 * 10 * 10 = 4,500 Mb (on 8 CD-ROM)

[Next: SMOS]

SMOS does not focus on

- LIMS and Database management
- MS identification of protein (database searching and pattern matching)

Statistical Model Of Spectra (SMOS)

- Scope
  - Mass spec proteomics
  - Other spectrum-based profiling methods, such as metabonomic
- Focus
  - Statistical modeling

Statistical Model Of Spectra (SMOS)

- Purpose
  - Standard for statistical data analysis, exchange, comparison, and verification
  - Audit trail for statistical manipulation of spectral data

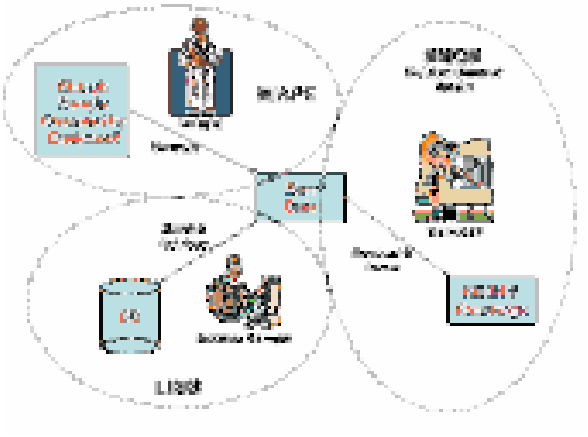
[Next: What is in SMOS]

Statistical Modeling of Spectra (SMOS)

- Single spectrum
  - Baseline removal, Smoothing etc
- A collection of spectrum
  - Normalization, Aggregation, Alignment etc.
- Raw spectrum -> Extracted Features
  - Peaks, Bins, Principle components
- Extracted Features -> Models
  - Clustering, Classification, and Survival
  - Biomarker discovery

[Next: graphic models of SMOS]

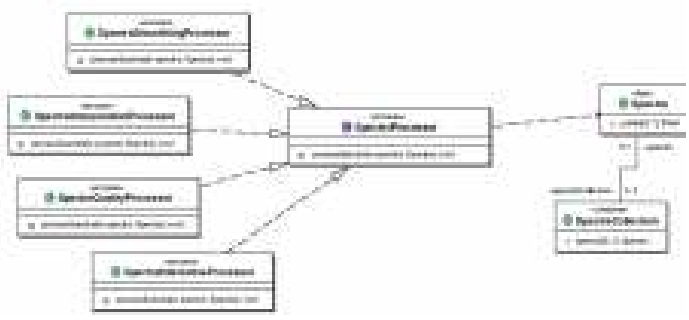
SMOS Integration Model



SMOS (part)

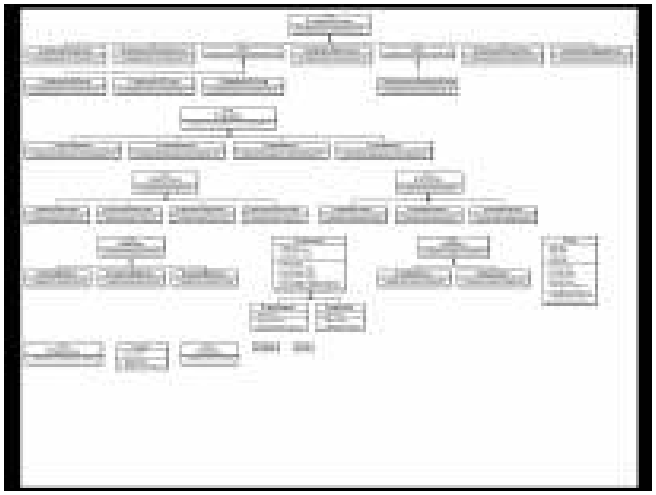
13 of 26

SMOS (part)



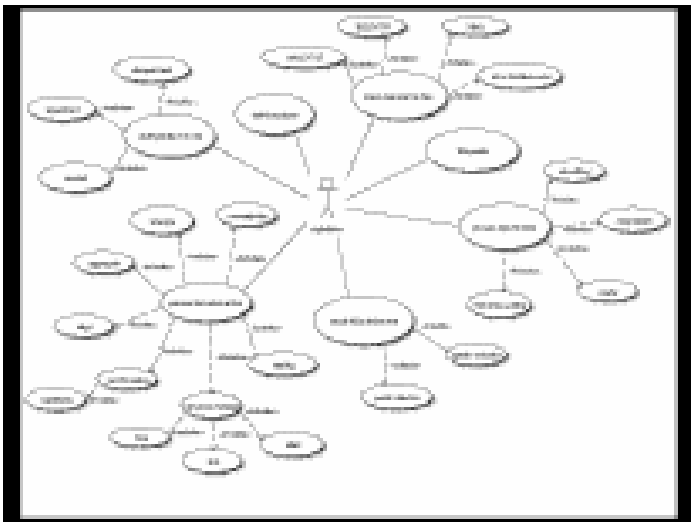
14 of 26

SMOS: UML model



15 of 26

SMOS: Use case Model



16 of 26

[next: use cases]

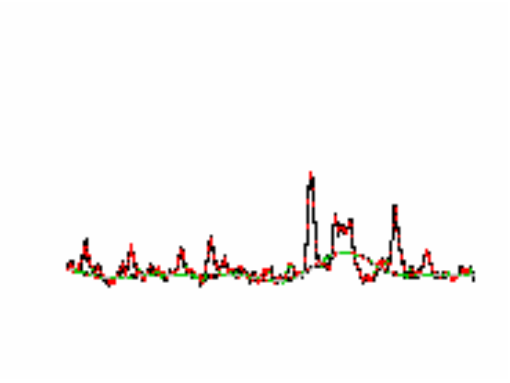
Use cases

- Use case: to get a better understanding of the problems and requirements in the scientific domain

  #1: Interactive browsing
  #2: Baseline removal
  #3: Data transfer

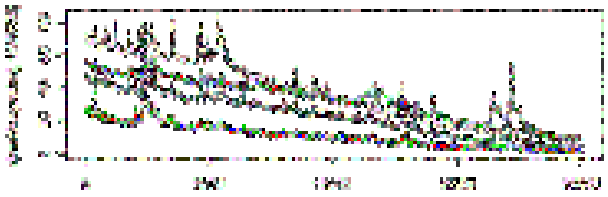
17 of 26

Use case #1: Interactive Browsing



18 of 26

Use case #2: baseline removal



19 of 26

Use Case #3: data transfer

- A biologist has profiling on 30 samples, generated 4,500 Mb of data.
- Wants to transfer the data over the internet to a statistician

20 of 26

Data Compression

- Faster transfer over the network
- Highly compressible
  - 36% of the original size
- Compression of scientific raw data
  - a common practice.
- Standards exist
  - netCDF
  - HDF5

[Next: engineering]

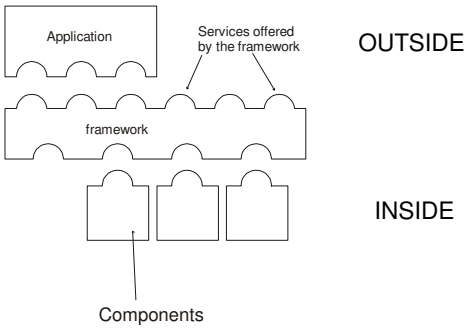
Software Engineering Aspect

- Reuse
- Modular
- V-model

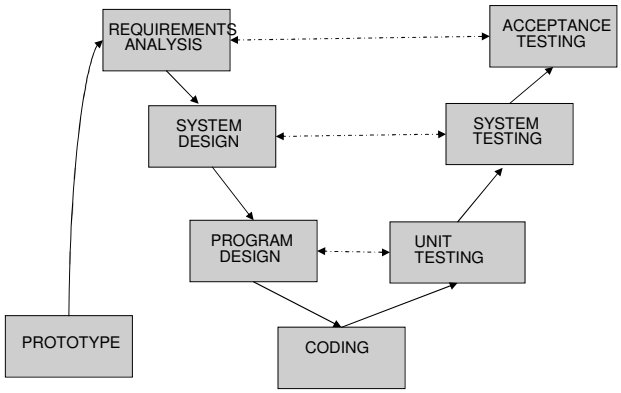
Building on the Experience of Others

- To avoid re-developing what already developed
  - Reuse of expertise
  - Reuse of standard designs and algorithms
  - Reuse of libraries built into languages
  - Reuse of frameworks

Modular Framework



V-model of Development



25 of 26

What's next

- SMOS: the model
  - Ontology model
  - UML model
- Rproteomics: the implementation
  - Requirement specification
  - Test data sets
  - Implementation and test

26 of 26

Acknowledgement

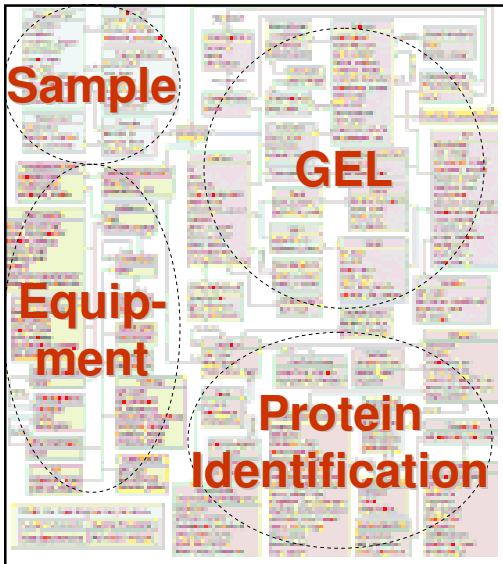
- Duke Radiology
  - Ned Patz
  - Mike Campa
- Duke Chemistry
  - Mike Fitzgerald

27 of 26

Standards: MIAPE and PSI

- Representation and archival of experimental method and data: MIAPE
  - Formerly known as PEDRo
  - Soon there will be PSI-ML and PSI-DB
  - Modular, extensible model
  - Nature Biotechnology 21: 247, 2003

28 of 26



MIAPE Model

29 of 26

Other Efforts

- Proteomics Database
  - Nature Biotechnology 22: 471, 2004
- Other efforts in the CaBIG Proteomics SIG
  - Fox Chase
  - Dartmouth
- MIAME: the microarray experience
  - Nature Genetics 29: 365, 2001

[The end]

30 of 26